University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2009 *Paper* 79

Composite likelihood Bayesian information criteria for model selection in high dimensional data

X Gao* Peter Xuekun Song†

*xingao@mathstat.yorku.ca

†University of Michigan, Ann Arbor, pxsong@umich.edu

Composite likelihood Bayesian information criteria for model selection in high dimensional data

X Gao and Peter Xuekun Song

Abstract

For high-dimensional data set with complicated dependency structures, the full likelihood approach often renders to intractable computational complexity. This imposes di±culty on model selection as most of the traditionally used information criteria require the evaluation of the full likelihood. We propose a composite likelihood version of the Bayesian information criterion (BIC) and establish its consistency property for the selection of the true underlying model. Under some mild regularity conditions, the proposed BIC is shown to be selection consistent, where the number of potential model parameters is allowed to increase to in¯nity at a certain rate of the sample size. Simulation studies demonstrate the empirical performance of this new BIC criterion, especially for the scenario that the number of parameters increases with the sample size.

Composite likelihood Bayesian information criteria for model selection in high dimensional data

By XIN GAO

Department of Mathematics and Statistics, York University, Toronto, Ontario Canada M3J 1P3

xingao@mathstat.yorku.ca

and PETER X.-K. SONG

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A.

pxsong@umich.edu

Summary

For high-dimensional data set with complicated dependency structures, the full likelihood approach often renders to intractable computational complexity. This imposes difficulty on model selection as most of the traditionally used information criteria require the evaluation of the full likelihood. We propose a composite likelihood version of the Bayesian information criterion (BIC) and establish its consistency property for the selection of the true underlying model. Under some mild regularity conditions, the proposed BIC is shown to be selection consistent, where the number of potential model parameters is allowed to increase to infinity at a certain rate of the sample size. Simulation studies demonstrate the empirical performance of this new BIC criterion, especially for the scenario that the number of parameters increases with the sample size.

Some key words: Consistency; Model selection; Pseudo-likelihood; Variable selection.

1. Introduction

In the analysis of high-dimensional data with complex dependency structures, the exact likelihood inference often renders to computational complexity. A compromise is to employ simpler pseudo-likelihoods, such as the composite likelihood approach (Lindsay, 1988 and Cox & Reid, 2004). A composite likelihood is constructed by low-dimensional likelihood objects defined over small subsets of data. This dimension reduction methodology on the likelihood function has been successfully applied in many areas, including for example, generalized linear mixed models (Renard et al., 2004), genetics (Fearnhead & Donnelly, 2002), spatial statistics (Hjort & Omre, 1994; Heagerty & Lele, 1998; Varin & Vidoni, 2005) and multivariate survival analysis (Parner, 2001 and Li & Lin, 2006). It has demonstrated to possess desirable theoretical properties, such as estimation consistency and asymptotic normality, and can be utilized to establish hypothesis testing procedures in a similar fashion to the classical likelihood ratio test; see a recent review paper by Varin (2008) and more references therein.

There often exist many potential candidate models to reveal the data generating mechanism. Model selection has become a very important issue in the endeavor of statistical modelling. In the work of Varin & Vidoni (2005), a composite likelihood information criterion analogous to Akaike's (1973) information criterion (AIC) has been proposed. Their method selects the model with the best prediction power by minimizing a composite Kullback-Leibler (KL) distance for a future experiment. The proposed first-order unbiased selection statistic contains two components: One is the composite loglikelihood of the data under a candidate model, and the other gives the penalty pertaining to the effective number of parameters in the model. In particular, when the composite likelihood takes the ordinary likelihood, the penalty term reduces to the exact number of parameters in the model, which coincides with the AIC. Note that AIC focuses on selecting models with best prediction power and that it is not a consistent model selection criterion (e.g. Haughton, 1988). As a result, AIC tends to favor over-fitting models. In effect, Varin & Vidoni's composite likelihood selection criterion resembles AIC, due to the fact that it penalizes the number of parameters at the rate of $O(1)$. In some applications, building a parsimonious model is critical to proper interpretations of, say, covariate effects; therefore, although over-fitting does not impact prediction much, it will be problematic in studies of association.

This paper focuses on the development of Bayesian information criterion (BIC) for the

composite likelihood methodology. BIC was first proposed by Schwarz (1978) in the paradigm of the maximum likelihood methodology. Later, many authors have extended it to other estimation methods, including Konish, Ando & Imoto (2004) in the penalized maximum likelihood method. See also Berger, Ghost & Mukhopadhyay (2003), Chakrabarti & Ghosh (2006) and Jiang (2007). Essentially, BIC penalizes more heavily on the number of parameters at the rate of $O(\log(n))$, and has been shown to be a consistent model selection criterion in many settings; for example, the linear model (Rao & Wu, 1989), the partially linear model (Wang, Li & Tsai, 2007), the change-point analysis (Yao, 1988; Csörgö & Horváth, 1997), and the longitudinal data analysis (Wang & Qu, 2009). Recently, Chen & Chen (2008) proposed an extended BIC (EBIC) criterion in the setting of linear regression models with high-dimensional covariates, where an extra penalty was proposed to penalize the dimension of model space that supposedly increases with the sample size. This penalty is essentially to enforce the selection of sparse models when the number of regression coefficients, P, tends to infinity as the sample size n increases. Such an EBIC criterion has shown to be a consistent model selection criterion in the case of linear models with large model spaces.

We consider a general statistical model for high-dimensional data with complicated correlation structures. One example of the high-dimensional data is correlated regression data with (e.g. longitudinal or clustered data), with a large number of covariates. When the composite likelihood is the method of parameter estimation, it is of interest to investigate whether BIC is available for model selection, and if so how it behaves. This motivates us to address the following three issues: (1) To define a BIC in the composite likelihood methodology, which will be referred to as the composite likelihood BIC or CL-BIC in the rest of this paper. This CL-BIC will be applicable for the situation where the number of parameters increases with the sample size. (2) To establish a large sample property of the model selection consistency for the proposed CL-BIC, which is a key advantage of BIC or its variants as seen in the literature. (3) To compare CL-BIC with Varin and Vidoni's composite likelihood AIC in order to understand the performances between AIC and BIC in the composite likelihood methodology. It is worth noting that Chen & Chen's EBIC becomes a special case of the proposed CL-BIC with the univariate composite likelihood in the linear model. In addition, the CL-BIC is also applicable for the full likelihood methodology, as the full likelihood is a special case of composite likelihood. In simulation studies given in Section 4, we include

a comparison of CL-BIC between the full and composite likelihood methods, since the full likelihood method serves as the gold standard in the simulation setting.

The paper is organized as follows. Section 2 presents the BIC in the composite likelihood framework, and Section 3 concentrates on the property of model selection consistency for the proposed CL-BIC. Section 4 illustrates the performance of the CL-BIC and comparisons with AIC via simulation studies, and Section 5 concludes the paper with some remarks. Some technical details are listed in the appendix.

2. Composite likelihood bayesian information criterion

2.1. Composite Likelihood

The composite likelihood (CL) paradigm (Lindsay, 1988) constitutes a rich class of pseudo-likelihoods based on marginal likelihood objects. Let $\{f(y;\psi), \psi \in \Psi\}$ be a parametric statistical model, with the parameter space $\Psi \subseteq \mathcal{R}^Q$. Let $Y = (Y_1', \ldots, Y_n')'$ denote the dataset, where $Y_i = (y_{i1}, \ldots, y_{im_i})'$ are the vector of observations sampled independently on unit i, $i = 1, \ldots, n$, from a study population. For convenience, we may regard the Y as the vectorized data, in which one observation y_{ij} is indexed by $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$. Since the methodology of composite likelihood lies in the idea of dimension reduction for the likelihood function, the parameter ψ would be partitioned as $\psi = (\theta, \eta)$, where θ is the parameter of interest to be estimated and η is the nuisance parameter that will not be estimated by the composite likelihood method. Consequently, the model selection with the composite likelihood methodology concerns with parameter θ, and the corresponding parameter space is $\Theta \subseteq \mathcal{R}^P$,, with dimension P possibly dependent on the sample size.

To form a composite likelihood, first consider a collection of index subsets $\mathcal{A} = \{A : A \subseteq \Omega\}$, where each element A is a subset of $\Omega = \{(i,j), j = 1, \ldots, m_i, i = 1, \ldots, n\}$. For a given unit i, similarly we denote $\mathcal{A}_i = \{A : A \subseteq \Omega_i\}$ with $\Omega_i = \{(i,j), j = 1, \ldots, m_i\}$. This implies that $\Omega = \cup_{i=1}^n \Omega_i$. Clearly the cardinality of set Ω, card(Ω), escalates as the sample size n increases. Then, let Y_A denotes the subset of the data with respect to set A, namely $Y_A = \{y_a, a \in A\}$. According to Lindsay (1988), a composite likelihood function is defined as

$$\mathrm{CL}(\theta; Y) = \prod_{A \in \mathcal{A}} L_A(\theta; Y)^{w_A} = \prod_{i=1}^{n} \prod_{A \in \mathcal{A}_i} L_A(\theta; Y)^{w_A}, \tag{1}$$

where $L_A(\theta; Y) = f(y_A; \theta)$ is the marginal likelihood with respect to composite set A, and $\{w_A\}$ is a set of suitable weights. It is easy to see that a singleton $\mathcal{A}_i = \{\Omega_i\}$ corresponds to the full likelihood, and $\mathcal{A}_i = \{\{1\}, \ldots, \{m_i\}\}$ gives rise to a composite likelihood of univariate margins. The log composite likelihood is $\mathrm{cl}(\theta; Y) = \sum_{i=1}^{n} \sum_{A \in \mathcal{A}_i} w_A \ell_A(\theta; Y)$, where $\mathrm{cl}(\theta; Y) = \log \mathrm{CL}(\theta; Y)$ and $\ell_A(\theta; Y) = \log L_A(\theta; Y)$. The maximum composite likelihood estimator (CLE) is given by

$$\hat{\theta}^c = \arg\max_{\theta \in \Theta} \mathrm{cl}(\theta; Y).$$

Since each term in (1) is a likelihood object, the resulting composite likelihood estimating equation $\nabla_\theta \mathrm{cl}(\theta; Y) = 0$ is unbiased under the assumption that these likelihood objects are valid marginal densities of the underlying joint parametric model $f(y; \psi)$. As usual, the composite likelihood estimate is obtained as a solution to this composite likelihood estimating equation. From the classical theory of estimating functions (e.g. Song, 2007, Chapter 3), the associated maximum composite likelihood estimator, $\hat{\theta}^c = \hat{\theta}^c(Y)$, is consistent and asymptotically normally distributed, under some mild regularity conditions. See also Varin (2008).

2.2. Bayes Information Criterion

Denote the true full parameter by $\psi_T = (\theta_T, \eta_T) \in \mathrm{int}(\Psi)$ and then the true marginal parameter by $\theta_T \in \mathrm{int}(\Theta)$. Consequently, the true full model is $f(y; \psi_T)$ and the true marginal model constitutes a set of true composite marginals $\{f(y_A; \theta_T), A \in \mathcal{A}_i\}$ for one unit, say i.

To derive BIC in the composite likelihood framework, we need some additional notations. Let $P = \dim(\Theta)$, and let s be a subset of $\{1, \ldots, P\}$. Denote by θ_s the parameter θ with those elements outside s being pre-specified as 0 or some known values. Because set s and a candidate marginal submodel $\{f(y_A; \theta_s), A \in \mathcal{A}\}$ correspond to each other uniquely, this submodel is simply denoted by s for convenience. Consequently, set $T \subseteq \{1, \ldots, P\}$ denotes the true marginal model.

Let d_s be the number of parameters under a marginal submodel s. Let $\mathcal{S}$ denote the model space of all possible submodels being considered. Associated with each submodel s, let $\pi_s(\theta_s|\omega_s)$ be the prior density of parameter θ_s, where ω_s is a certain given hyper-parameter, and let $p(s)$ be the prior probability of the occurrence of the submodel defined on space $\mathcal{S}$.

Following Schwarz's (1978) BIC method, we would select the best submodel that maximizes the posterior probability of the submodel given the equal priors. When the full likelihood is numerically prohibitive to compute, we replace the full likelihood by the composite likelihood defined in (1). This leads to a modified posterior probability, referred to as the *composite posterior probability* (CPP), for a submodel s,

$$\mathrm{CPP}(s|Y) = \frac{p(s)\int_{\Theta_s}\{\mathrm{CL}(\theta_s;Y)\}\pi_s(\theta_s|\omega_s)d\theta_s}{\sum_{r\in\mathcal{S}}p(r)\int_{\Theta_r}\{\mathrm{CL}(\theta_r;Y)\}\pi_r(\theta_r|\omega_r)d\theta_r},\ s\in\mathcal{S}. \tag{2}$$

It is worth pointing out that although $\mathrm{CL}(\theta_s;Y)$ in (1) is generally not a proper probability distribution, this CPP defined in (2) is due to the normalization, and hence it has the similar interpretation as the original BIC; that is, the best model is selected according to the maximum composite posterior probability among all the possible submodels.

Clearly, maximizing the composite posterior in (2) is equivalent to maximizing its numerator, $p(s)\int_{\Theta_s}\{\mathrm{CL}(\theta_s;Y)\}\pi_s(\theta_s|\omega_s)d\theta_s$, because the denominator is model free. Under certain regularity conditions (*e.g.* Tierney & Kadane, 1986; Tierney, Kass & Kadane, 1989), we approximate the integral through the Laplace formula, given as follows:

$$\int_{\Theta_s}\{\mathrm{CL}(\theta_s;Y)\}\pi_s(\theta_s|\omega_s)d\theta_s = \frac{(2\pi)^{\frac{d_s}{2}}}{n^{\frac{d_s}{2}}|Q_s(\tilde{\theta}_s)|^{\frac{1}{2}}}\exp\left\{nq_s(\tilde{\theta}_s|Y,\omega_s)\right\}\{1+O_p(n^{-1})\},$$

where $q_s(\tilde{\theta}_s|Y,\omega_s) = \frac{1}{n}\log\{\mathrm{CL}(\theta_s;Y)\pi_s(\theta_s|\omega_s)\}|_{\tilde{\theta}_s}$, and $Q_s(\tilde{\theta}_s) = \frac{\partial^2 q_s(\theta_s|Y,\omega_s)}{\partial\theta_s\partial\theta_s'}|_{\tilde{\theta}_s}$, with $\tilde{\theta}_s$ being the mode of $q_s(\theta_s|Y,\omega_s)$. It follows that

$$\begin{aligned}-2\log\left[p(s)\int\{\mathrm{CL}(\theta_s;Y)\}\pi_s(\theta_s|\omega_s)d\theta_s\right] &= -2\log\mathrm{CL}(\tilde{\theta}_s;Y) - 2\log\pi_s(\tilde{\theta}_s|\omega_s) + d_s\log(n)\\ &\quad +\log|Q_s(\tilde{\theta}_s)| - 2\log p(s) - d_s\log(2\pi)\\ &\quad +O_p(n^{-1}).\end{aligned} \tag{3}$$

Under the assumption that $\log\{\pi_s(\theta_s|\omega_s)\} = O(1)$, the mode $\tilde{\theta}_s$ of $q_s(\theta_s|Y,\omega_s)$ satisfies the following asymptotic expansion:

$$\tilde{\theta}_s = \hat{\theta}_s^c + \frac{1}{n}\left\{J_s(\hat{\theta}_s^c)\right\}^{-1}\left\{\frac{\partial}{\partial\theta_s}\log\pi_s(\theta_s|\omega_s)\Big|_{\hat{\theta}_s^c}\right\} + O_p(n^{-2}),$$

where $\hat{\theta}_s^c$ is the maximum composite likelihood estimator of θ_s under submodel s and

$$J_s(\hat{\theta}_s^c) = -\frac{1}{n}\frac{\partial^2\log\mathrm{CL}(\theta_s;Y)}{\partial\theta_s\partial\theta_s'}\Big|_{\hat{\theta}_s^c}. \tag{4}$$

Furthermore, when the prior $\pi_s(\theta_s|\omega_s)$ is chosen to be sufficiently flat in the neighborhood of the $\hat{\theta}_s^c$, retaining the leading terms in the approximation (3) at the rate of $O_p(1)$ or higher results in

$$\mathrm{BIC}(s) = -2\log \mathrm{CL}(\hat{\theta}_s^c; Y) + d_s \log(n) - 2\log\{p(s)\}, \tag{5}$$

where the last term may diverge at a faster rate than $O_p(1)$ in the case where the number of the parameters increases to infinity with the increase of sample size.

In the case of the composite likelihood AIC, because the composite likelihood is a kind of pseudolikelihood, Varin & Vidoni (2005) reached the form for the effective number of degrees of freedom, namely $d_s^* = \mathrm{trace}\,(H_s^{-1}V_s)$, where the sensitivity and variability matrices are given by, respectively,

$$H_s = \mathrm{E}_{\psi_{T,0}}\left\{-\frac{\partial^2 \log \mathrm{CL}(\theta_s; Y)}{\partial\theta_s \partial\theta_s'}\right\}, \text{ and } V_s = \mathrm{var}_{\psi_{T,0}}\left\{\frac{\partial \log \mathrm{CL}(\theta_s; Y)}{\partial\theta_s}\right\}, \tag{6}$$

where the expectations are taken under the true distribution of data generation, $f(y; \psi_{T,0})$.

Using d_s^* as an indicator of model complexity has been widely accepted (*e.g.* Pan, 2001) in the pseudo-likelihood methodology. Thus, we adopt this d_s^* into our BIC in the rest of this paper, and a consistent estimator is denoted by $\hat{d}_s^* = \mathrm{trace}\,(\hat{H}_s^{-1}\hat{V}_s)$, which will be used in all the related computation. It turns out that such modification is necessary to ensure the selection consistency, as shown in Section 3. With regard to consistent estimation of both $\hat{H}_s$ and $\hat{V}_s$, without the estimation of the nuisance parameter η, readers may refer to Varin & Vidoni (2005) for detail.

2.3. Sparsity via Penalization

In the conventional setting where the number of parameters P is fixed (or not dependent on the sample size n), it is commonly assumed that each submodel s has an equal probability of being selected, namely a uniform prior over the model space, $p(s) = 1/\mathrm{card}(\mathcal{S})$, where $\mathrm{card}(\mathcal{S})$ is the cardinality of $\mathcal{S}$. Consequently, the last term in the BIC (5), $-2\log\{p(s)\}$, is of order $O_p(1)$, and hence can be removed from the expression. As a result, (5) reduces to the classical Schwarz's BIC.

A much more challenging task of model selection in the high-dimensional data analysis is that P is not fixed and increases as the sample size rises. Suppose that $P_n = O(n^\kappa)$, with $\kappa > 0$. In this case, the equal probability prior will actually favor models with more parameters;

see for example Chen and Chen (2008). In many practical studies, important attributes are typically handful, in spite of a large P_n. This naturally necessitates the imposition of lower preferences on models with a large number of parameters; in other words, an additional penalty is required in BIC to ensure an increasing chance of selecting models with sparsity. This can be done by assigning priors through a stratified sampling scheme proposed by Chen and Chen (2008). To proceed, first partition the model space into submodel spaces $\mathcal{S} = \cup_{k=1}^{P_n} \mathcal{S}_k$, where each $\mathcal{S}_k$ contains models with k parameters. For example, $\mathcal{S}_1$ is a collection of all the models containing one parameter. Let $\tau(\mathcal{S}_k) = \text{card}(\mathcal{S}_k)$ be the size of $\mathcal{S}_k$. Obviously, $\tau(\mathcal{S}_1) = P_n$. Within a given subspace $\mathcal{S}_k$, an equal probability prior is imposed as $p(s|\mathcal{S}_k) = 1/\tau(\mathcal{S}_k), s \in \mathcal{S}_k$. Obviously, smaller models gain higher prior probabilities. Moreover, specifying prior probabilities for these subspaces proportional to their sizes, say $p(\mathcal{S}_k) \propto \{\tau(\mathcal{S}_k)\}^{\xi}$ for some $\xi \in [0, 1]$, we obtain that the prior probability of a submodel s being selected via this stratified sampling procedure is proportional to $\tau(\mathcal{S}_k)^{-\gamma}$, with $\gamma = 1 - \xi \in [0, 1]$. In particular, when $\xi = 1$, and $\gamma = 0$, this stratified prior reduces to the unstratified uniform prior considered in the conventional BIC case.

Consequently, we reach a composite likelihood BIC for model selection given as follows:

$$\text{CL-BIC}(s) = -2 \log \text{CL}(\hat{\theta}_s^c; Y) + \hat{d}_s^* \log(n) + 2\gamma \log\{\tau(\mathcal{S}_{\hat{d}_s^*})\}. \tag{7}$$

In (7), the first term is minus twice of the composite loglikelihood that reflects the goodness-of-fit for a given submodel s, the second term is the penalty for the model complexity; and the third term is the penalty for the enforcement of sparsity on the model selected. The coefficient γ tunes the degree of preference on large sized models. The larger the γ, the more favorable a sparse model. It is worth noting that although the above development of the CL-BIC is derived by a Bayesian approach, the practical use of the CL-BIC will not require any specific Bayesian model components. For example, no prior distributions are needed in the evaluation of the CL-BIC (7).

3. Selection consistency

Given an arbitrary submodel s, it may be (1) the true marginal model T, with the parameter vector θ_T that contains d_T components; or (2) an under-fitting model $s-$, which is a misspecified model under which the model parameter θ_T is not a subset of the θ_{s-}, i.e.

$\theta_{s-} \not\supseteq \theta_T$; or (3) an over-fitting model $s+$, in that the parameter vector θ_{s+} contains the true θ_T but is not identical to the θ_T, that is, $\theta_T \subset \theta_{s+}$ and $\theta_{s+} \neq \theta_T$. For these three scenarios, let CL-BIC$(s), s = T, s-, s+$ denote the composite likelihood BIC criteria obtained under the true (T), under-fitting ($s-$) and over-fitting marginal models ($s+$), respectively.

In this paper, we assume the conventional regularity conditions required for consistency and asymptotic normality of the maximum likelihood estimator (Cox and Hinkley, 1974). Furthermore, we assume four additional regularity conditions needed by the composite likelihood estimation in connection to model misspecification (*e.g.* White, 1982; Varin & Vidoni, 2005), detailed as follows.

Assumption 1 (A1). For each submodel s, the parameter space Θ_s is a compact subset of $\mathcal{R}^{d_s}$, and for fixed Y, $\mathrm{cl}(\theta_s; Y)$ is twice continuously differentiable with respect to θ_s.

Assumption 2 (A2). (a) For each submodel s, $|\mathrm{cl}(\theta_s; Y)|$, $|\partial \mathrm{cl}(\theta_s; Y)/\partial\theta_{s_i} \cdot \partial \mathrm{cl}(\theta_s; Y)/\partial\theta_{s_j}|$, $|\partial^2 \mathrm{cl}(\theta_s; Y)/\partial\theta_{s_i}\theta_{s_j}|, i, j = 1, \ldots, d_s$, are dominated by functions integrable with respect to the probability measure of the true marginal model for all $\theta_s \in \Theta_s$. (b) Denote the log composite likelihood ratio (CLR) between two marginal submodels s and s' by

$$\lambda_{s'|s}(Y; \theta_{s'}, \theta_s) = \log\left\{\frac{\mathrm{CL}(\theta_{s'}; Y)}{\mathrm{CL}(\theta_s; Y)}\right\} = \mathrm{cl}(\theta_{s'}; Y) - \mathrm{cl}(\theta_s; Y). \tag{8}$$

Assume $\mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y; \theta_{T,0}, \theta_s)\}$ exists for all θ_s, and has a unique minimum at $\theta_{s,0} \in \mathrm{int}(\Theta_s)$. Here $\psi_{T,0}$ is the true value of the parameter ψ_T under the true full model $f(y; \psi)$.

It is easy to see that this $\theta_{s,0}$ effectively defines the pseudo true value of parameter θ_s in Θ_s under a misspecified model s, which minimizes the expected composite Kullback-Leibler distance (Varin & Vidoni, 2005) between the true marginal model and a marginal submodel s. That is, $\theta_{s,0} = \arg\min_{\theta_s \in \Theta_s} \mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y; \theta_{T,0}, \theta_s)\}$.

Assumption 3 (A3). The composite likelihood estimator $\hat{\theta}_s^c$ is consistent, $\hat{\theta}_s^c \xrightarrow{p} \theta_{s,0}$, and asymptotically normally distributed, $\sqrt{n}(\hat{\theta}_s^c - \theta_{s,0}) \xrightarrow{d} N_{d_s}(0, G^{-1})$, where G is the Godambe information matrix (or the sandwich covariance).

To establish the model selection consistency of the CL-BIC in the case of large P and small n, namely $P = P_n = O(n^\kappa)$, as $n \to \infty$ for some $\kappa > 0$, assumption 4 below is imposed to ensure that the true marginal model T is asymptotically identifiable in the large model space.

Assumption 4 (A4). (a) $\mathrm{var}_{\psi_{T,0}}\{\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0})\}$ exists; (b) $\max_{s \in \mathcal{S}} d_s^*/d_s \leq C_0$ for some constant C_0, where d_s^* is the effective number of degrees of freedom; and (c) both of the

following conditions hold:

$$\lim_{n\to\infty}\min_{s\in\mathcal{S}}\left\{(\log n)^{-\frac{1}{2}}\frac{\mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y;\theta_{T,0},\theta_{s,0})\}}{[\mathrm{var}_{\psi_{T,0}}\{\lambda_{T|s}(Y;\theta_{T,0},\theta_{s,0})\}]^{\frac{1}{2}}}, s\neq T, d_s\leq d_T\right\}=\infty, \tag{9}$$

$$\liminf_{n\to\infty}\min_{s\in\mathcal{S}}\left\{(\log n)^{-\frac{1}{2}}[\mathrm{var}_{\psi_{T,0}}\{\lambda_{T|s}(Y;\theta_{T,0},\theta_{s,0})\}]^{\frac{1}{2}}, s\neq T, d_s\leq d_T\right\}\geq C_1, \tag{10}$$

for a positive constant C_1.

In effect, assumption 4 implies that for each under-fitting marginal submodel,

$$\mathrm{var}_{\psi_{T,0}}\{\lambda_{T|s}(Y;\theta_{T,0},\theta_{s,0})\}=o_p\left((\log n)^{-1/2}\mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y;\theta_{T,0},\theta_{s,0})\}\right), \tag{11}$$

$$\lim_{n\to\infty}(\log n)^{-1}\mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y;\theta_{T,0},\theta_{s,0})\}=\infty. \tag{12}$$

Assumption 4 is a generalization of the asymptotical identifiability condition given by Chen and Chen (2008) in the linear model setting. Proposition 1 below shows this claim. To proceed, consider a linear model, $Y=X\theta+\epsilon$, where $\epsilon\sim N_n(0,\sigma^2 I)$. Then, X_T and X_s denote the design matrices of the true model and a candidate model, respectively, with respective vectors of the regression coefficients, θ_T and θ_s. Moreover, the true null value is $\theta_{T,0}$ under the true model, and the pseudo null value is $\theta_{s,0}$ under the candidate model.

Proposition 1. *In the linear model, assumption A4(c) given in (9) and (10) reduces to the following condition:*

$$\lim_{n\to\infty}\min_{s\in\mathcal{S}}\left\{(\log n)^{-1}\Delta_n(s), s\neq T, d_s\leq d_T\right\}=\infty, \tag{13}$$

with $\Delta_n(s)=||X_T\theta_T-D(s)X_T\theta_{T,0}||_2^2$, *and* $D(s)=X_s(X_s'X_s)^{-1}X_s'$ *is the hat matrix.*

The proof of Proposition 1 is presented in the appendix.

We establish the model selection consistency of the CL-BIC in the following two theorems. The first concerns with the consistency for the under-fitting models and the second for the over-fitting models.

Theorem 2. *Suppose the number of parameters* $P_n=O(n^\kappa), \kappa>0$. *Under the regularity conditions (A1)-(A4), for any* $\gamma>0$,

$$P_{\psi_{T,0}}\{\textit{CL-BIC}(T)<\textit{CL-BIC}(s-)\}\to 1,\ \textit{as } n\to\infty.$$

To prove Theorem 2, we need the following three lemmas.

Lemma 3. *Given a candidate marginal submodel $s \in \mathcal{S}_{d_s}$ (either under- or over-fitting), consider a quadratic form:*

$$Q(s) = \left(\hat{\theta}_s^c - \theta_{s,0}\right)' \left\{ \frac{-\partial^2 cl(\theta_s; Y)}{\partial \theta_s \partial \theta_s'} \right\} \Big|_{\hat{\theta}_s^c} \left(\hat{\theta}_s^c - \theta_{s,0}\right), \ s \in \mathcal{S}_{d_s}. \tag{14}$$

Then as $n \to \infty$, $Q(s)$ asymptotically follows a mixture of chi-square distributions, $\sum_{j=1}^{d_s} \zeta_j(s)\chi_j^2$, where $\zeta_1(s), \ldots, \zeta_{d_s}(s)$ are the eigenvalues of the matrix $H_s^{-1}V_s$, and $\chi_j^2, j = 1, \ldots, d_s$ are i.i.d. according to the chi-square distribution with 1 degree of freedom. Both H_s and V_s are given in (6).

Lemma 4. *For a given set of K independent standard normal random variables $Z_1, \ldots, Z_K$,*

$$\max\{Z_i, i = 1, \ldots, K\} = O_p(\sqrt{\log K}).$$

Lemma 5. *For the quadratic form $Q(s)$ given in (14), under the same setting of Lemma 3,*

$$\max\{Q(s), s \in \mathcal{S}_{d_s}\} = O_p(d_s^* \log(P_n)),$$

where $d_s^* = \sum_{j=1}^{d_s} \zeta_j(s)$, with eigenvalues $\zeta_j(s)$'s being given in Lemma 3.

The proofs of the three lemmas are detailed in the appendix. Now we prove Theorem 2.

Proof. In the following, we use s, instead of $s-$, to denote an under-fitting submodel just for the convenience of exposition. Expanding the composite log-likelihood around the composite likelihood estimate $\hat{\theta}_s^c$, we obtain

$$\begin{aligned} \mathrm{cl}(\theta_{s,0}; Y) &= \mathrm{cl}(\hat{\theta}_s^c; Y) + \frac{\partial \mathrm{cl}(\theta_s; Y)}{\partial \theta_s}|_{\hat{\theta}_s^c}(\theta_{s,0} - \hat{\theta}_s^c) + \\ &\quad \frac{1}{2}(\theta_{s,0} - \hat{\theta}_s^c)' \frac{\partial^2 \mathrm{cl}(\theta_s; Y)}{\partial \theta_s \partial \theta_s'}|_{\hat{\theta}_s^c}(\theta_{s,0} - \hat{\theta}_s^c)\{1 + o_p(1)\}, \end{aligned} \tag{15}$$

where the second term in (15) is zero, and the third term is $-\frac{1}{2}Q(s)$. It follows from Lemma 3 that for a large n,

$$-2\{\mathrm{cl}(\theta_{s,0}; Y) - \mathrm{cl}(\hat{\theta}_s^c; Y)\} = Q(s) + o_p(1) = \sum_{j=1}^{d_s} \zeta_j \chi_j^2 + o_p(1),$$

where both ζ_j and χ_j^2 are given in Lemma 3. Thus, by Lemma 5 for a submodel $s \in \mathcal{S}_{d_s}$, we obtain

$$\max_{s \in \mathcal{S}_{d_s}} 2\{\mathrm{cl}(\hat{\theta}_s^c; Y) - \mathrm{cl}(\theta_{s,0}; Y)\} = O_p(\log n), \tag{16}$$

where $P_n = O(n^\kappa)$. On the other hand, for the true marginal model T, Lemma 3 implies that

$$2\{\mathrm{cl}(\hat{\theta}_T^c; Y) - \mathrm{cl}(\theta_{T,0}; Y)\} = \sum_{j=1}^{d_T} \zeta_j(T)\chi_j^2 = O_p(1), \tag{17}$$

where $\zeta_j(T), j = 1, \ldots, d_T$ are the eigenvalues of $H_T^{-1}V_T$.

Furthermore, by the Central Limit Theorem, for each submodel s,

$$Z(s) = \frac{\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0}) - \mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0})\}}{\mathrm{var}_{\psi_{T,0}}\{\lambda_{T|s-}(Y; \theta_{T,0}, \theta_{s-,0})\}} \xrightarrow{d} N(0,1).$$

Thus, Lemma 4 and equation (11) entail that

$$\begin{aligned}
&\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0}) - \mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s-}(Y; \theta_{T,0}, \theta_{s,0})\} \\
&\leq \max\{Z(s), s \in \mathcal{S}_{d_s}\} \left[\mathrm{var}_{\psi_{T,0}}\{\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0})\}\right]^{\frac{1}{2}} \{1 + o_p(1)\} \\
&= O_p(\sqrt{\log n}) \left[\mathrm{var}_{\psi_{T,0}}\{\lambda_{T|s-}(Y; \theta_{T,0}, \theta_{s-,0})\}\right]^{\frac{1}{2}} \{1 + o_p(1)\} \\
&= o_p(\mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s-}(Y; \theta_{T,0}, \theta_{s-,0})\}),
\end{aligned}$$

Summarizing all the results, we have for an under-fitting submodel s,

$$\begin{aligned}
&-2\{\mathrm{cl}(\hat{\theta}_s^c; Y) - \mathrm{cl}(\hat{\theta}_T^c; Y)\} \\
&= -2\{\mathrm{cl}(\hat{\theta}_s^c; Y) - \mathrm{cl}(\theta_{s,0}; Y)\} + 2\{\mathrm{cl}(\hat{\theta}_T^c; Y) - \mathrm{cl}(\theta_{T,0}; Y)\} \\
&\quad +2[\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0}) - \mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0})\}] + 2\mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0})\} \\
&= \mathrm{E}_{\psi_{T,0}}\{\lambda_{T|s}(Y; \theta_{T,0}, \theta_{s,0})\}\{1 + o_p(1)\}.
\end{aligned}$$

It follows from equation (12) that $\lim_{n\to\infty} -2\{(\log n)^{-1}(\mathrm{cl}(\hat{\theta}_s^c; Y) - \mathrm{cl}(\hat{\theta}_T^c; Y))\} = \infty$. Since the penalty terms in the difference of $\{\text{CL-BIC}(s-) - \text{CL-BIC}(T)\}$ are both of order $O_p(\log n)$, we obtain $P_{\psi_{T,0}}\{\text{CL-BIC}(T) < \text{CL-BIC}(s-)\} \to 1$. □

Next we consider the over-fitting scenario.

Theorem 6. *Suppose the number of parameters $P_n = O(n^\kappa), \kappa > 0$, and suppose the regularity conditions A1-A4 hold. When $\gamma > 1 - 1/(2\kappa)$, for an over-fitting model $s+$,*

$$P_{\psi_{T,0}}\{\textit{CL-BIC}(T) < \textit{CL-BIC}(s+)\} \to 1, \ \textit{as } n \to \infty.$$

Proof. Without the loss of generality, for an over-fitting marginal submodel $s+$, we write $\theta_{s+} = (\theta_T, \theta_W)$, where θ_W denotes the vector of nuisance parameters. Accordingly, $\hat{\theta}_{s+}^c =$

$(\hat{\theta}_T^c, \hat{\theta}_W^c)$ denote the maximum composite likelihood estimator obtained from the over-fitting model $s+$ with the corresponding composite log-likelihood denoted as $\mathrm{cl}_+(\hat{\theta}_T^c, \hat{\theta}_W^c; Y)$. Likewise, let $\tilde{\theta}_T^c$ be the maximum composite likelihood estimator obtained from the true marginal submodel T with the corresponding composite log-likelihood denoted as $\mathrm{cl}_T(\tilde{\theta}_T^c, 0; Y)$.

Applying Taylor expansion on $\mathrm{cl}_T(\hat{\theta}_T^c, 0; Y)$ around $\tilde{\theta}_T^c$, we obtain

$$\begin{aligned}&\mathrm{cl}_T(\hat{\theta}_T^c, 0; Y) - \mathrm{cl}_T(\tilde{\theta}_T^c, 0; Y) = \\ &\quad \frac{1}{2}(\hat{\theta}_T^c - \tilde{\theta}_T^c)' \left\{ \frac{\partial^2 \mathrm{cl}_T(\theta_T, \theta_W; Y)}{\partial \theta_T \partial \theta_T'} \Big|_{(\tilde{\theta}_T^c, 0)} \right\} (\hat{\theta}_T^c - \tilde{\theta}_T^c)\{1 + o_p(1)\}. \end{aligned} \tag{18}$$

Given that both $\tilde{\theta}_T^c$ and $\hat{\theta}_T^c$ are root-n consistent for the true null value $\theta_{T,0}$, we have $(\hat{\theta}_T^c - \tilde{\theta}_T^c) = O_p(n^{-\frac{1}{2}})$. Then, similar to the proof of Lemma 3 in the appendix, it is easy to show that the quadratic form in (18) is of order $O_p(1)$.

On the other hand, applying Taylor expansion on $\mathrm{cl}_+(\theta_T, \theta_W)$ around $\hat{\theta}_W^c$ and then evaluating the expansion at $\hat{\theta}_T = \hat{\theta}_T^c$ and $\theta_W = 0$, we obtain

$$\mathrm{cl}_+(\hat{\theta}_T^c, 0; Y) - \mathrm{cl}_+(\hat{\theta}_T^c, \hat{\theta}_W^c; Y) = \frac{1}{2}(\hat{\theta}_W^c)' \left\{ \frac{\partial^2 \mathrm{cl}_+(\hat{\theta}_T^c, \theta_W; Y)}{\partial \theta_W \partial \theta_W'} \Big|_{\hat{\theta}_W} \right\} \hat{\theta}_W^c \{1 + o_p(1)\}. \tag{19}$$

It is known that under the over-fitting model, $\hat{\theta}_W^c$ is consistent and asymptotically normally distributed, $N(0, H_W^{-1} V_W H_W^{-1})$, where

$$H_W = \mathrm{E}_{\psi_{T,0}} \left\{ -\frac{\partial^2 \mathrm{cl}_+(\theta_T, \theta_W; Y)}{\partial \theta_W \partial \theta_W^T} \right\}, \text{ and } V_W = \mathrm{var}_{\psi_{T,0}} \left\{ \frac{\partial \mathrm{cl}_+(\theta_T, \theta_W; Y)}{\partial \theta_W} \right\}.$$

Let $\zeta_1(W), \ldots, \zeta_r(W)$ be the r eigenvalues of matrix $H_W^{-1} V_W$, where $r = d_{s+} - d_T$. According to Lemma 3, $-2\{\mathrm{cl}_+(\hat{\theta}_T^c, \hat{\theta}_W^c; Y) - \mathrm{cl}_+(\hat{\theta}_T^c, 0; Y)\}$ follows asymptotically a mixture of chi-square distributions, $\sum_{j=1}^r \zeta_j(W) \chi_j^2$. Then, Lemma 5 implies that

$$\max \left[-2\{\mathrm{cl}_+(\hat{\theta}_T^c, \hat{\theta}_W^c; Y) - \mathrm{cl}_+(\hat{\theta}_T^c, 0; Y)\}, s+ \in \mathcal{S}_{d_{s+}} \right] = O_p((d_s^* - d_T^*) \log(P_n)). \tag{20}$$

Finally, the difference of the CL-BIC criteria between the over-fitting and true marginal submodels is given by

$$\begin{aligned}&\text{CL-BIC}(s+) - \text{CL-BIC}(T) \\ &= -2 \left\{ \mathrm{cl}_+ \left(\hat{\theta}_{s+}^c; Y \right) - \mathrm{cl}_T \left(\tilde{\theta}_T^c; Y \right) \right\} + (d_s^* - d_T^*) \log(n) + 2\gamma \left\{ \log(\tau(\mathcal{S}_{d_{s+}^*})) - \log(\tau(\mathcal{S}_{d_T^*})) \right\} \\ &\geq -2 \left\{ \mathrm{cl}_+ \left(\hat{\theta}_{s+}^c; Y \right) - \mathrm{cl}_+ \left(\hat{\theta}_T^c; Y \right) \right\} - 2 \left\{ \mathrm{cl}_+ \left(\hat{\theta}_T^c; Y \right) - \mathrm{cl}_T \left(\hat{\theta}_T^c; Y \right) \right\} \\ &\quad -2 \left\{ \mathrm{cl}_T \left(\hat{\theta}_T^c; Y \right) - \mathrm{cl}_T \left(\tilde{\theta}_T^c; Y \right) \right\} + (d_s^* - d_T^*) \log(n) + 2\gamma (d_s^* - d_T^*) \log(P_n) \\ &\geq (d_s^* - d_T^*) O_p(\log(n) + 2(\gamma - 1) \log(P_n)). \end{aligned}$$

When $\gamma \geq 1 - 1/(2\kappa)$, we have $P_{\psi_{T,0}}\{\text{CL-EBIC}(T) < \text{CL-BIC}(s+)\} \to 1$, as $n \to \infty$, given $P_n = O(n^\kappa), \kappa > 0$. □

4. Simulation

To examine the performance of the CL-BIC, we conduct two Monte Carlo simulation experiments that primarily concern the selection of tuning parameter in the LASSO approach (Tibshirani, 1996). The CL-BIC assists us to select an optimal tuning parameter by comparing the CL-BIC values among a set of sequentially generated candidate models. We consider the multivariate familial data analysis discussed in Zhao and Joe (2005). The sample is drawn from families with inter-correlations among individuals in a family. Denote the number of families by n and the number of members in each family by m. The response vector of measurements for the i-th family is denoted by $Y_i = (y_{i1}, \ldots, y_{im})'$. Associated is a set of covariates at the individual level, $X_i = (x_{i1}, \ldots, x_{im})'$, with $x_{ik} = (x_{ik1}, \ldots, x_{ikP})'$, representing the P covariates observed for the k-th individual in the i-th family. The first study concerns a multivariate normal model, in which Y_i follows a multivariate normal distribution, $N_m(\mu_i, \Sigma)$, where the mean vector is governed by a linear model, $\mu_i = X_i\beta$, with $\beta = (\beta_1, \ldots, \beta_P)$. The covariance matrix Σ was specified according to an exchangeable dependence structure, $\sigma_{k,k'} = \rho$. The second simulation study is based on a multivariate probit model, in which the binary response vector arises from a dichotomization of an underlying multivariate normal of exactly the same specification as given in the first study. In addition, among all the covariate coefficients, most of them are zero while a small subset are non-zero. We wish to select the significant covariates among the P candidates.

4.1. Multivariate normal model

We consider two different scenarios. In the first scenario, we set $P = 30$, $n = 200$ and $m = 4$. The covariates are generated from a multivariate normal with the standard normal $N(0, 1)$ marginals and inter-correlation $\text{Cov}(x_{ikp}, x_{ikp'}) = 0.2$. The within-family correlation ρ is set to either 0.3 or 0.6. The regression coefficients of the true marginal model are set to two values $\beta_1(T) = (0.1, 0.2, 0.4, 0.1, 0.4, 0.2, 0.3, 0.4, 0.5, 0.3)$, or $\beta_2(T) = (0.5, 0.1, 0.4, 0.3, 0.5, 0.1, 0.004, 0.04, 0.03,\ 0.003)$, with the other 20 coefficients set to zero.

The second case contains 10 nonzero coefficients, while four of them are too small and considered to be no useful and not be used to compute the positive selection rates. This setup can help us evaluate the performance of the model selection criteria when the covariates have different levels of effect. In the second scenario, we set $P = 1000$, $n = 200$, and $m = 4$. The covariates are partitioned into 20 blocks of 50 each. Within each block, the covariates are generated from a multivariate normal with univariate standard normal marginals and equal inter-covariate correlation 0.2, and covariates from different blocks have zero correlations. Similarly, the within-family correlation ρ is set to either 0.3 or 0.6. The regression coefficients of the true marginal model are set to the same values as those in scenario I with the other 990 coefficients set to zero.

To apply LASSO, we impose penalization on the composite likelihood with L_1 penalty. We gradually increase the tuning parameter in the penalty term and obtain a sequence of nested models. Under scenario II with $P >> n$, we randomly partition the 1000 covariates into 8 disjoint subsets of 125 covariates each and apply the penalized composite likelihood on each subset. We then pool the reduced subsets of covariates together and perform the same procedure to obtain the sequence of nested models, at which the CL-BIC is computed to determine the optimal tuning.

For each candidate model, the CL-BIC is evaluated under either the univariate (or onewise) composite likelihood $\sum_{i=1}^{n}\sum_{k=1}^{m} \mathrm{cl}(y_{ik};\beta)$, or the pairwise composite loglikelihood $\sum_{i=1}^{n}\sum_{k<k'} \mathrm{cl}(y_{ik}, y_{ik'};\beta)$. The resulting two CL-BIC criteria are denoted as CL_U-BIC or CL_B-BIC, respectively. For the purpose of comparison, we also include Varin and Vidoni's CL_U-AIC based on the univariate composite likelihood, and Chen and Chen's EBIC based on the full likelihood that serves as the gold standard. The two versions of CL-BIC and the EBIC are calculated with $\gamma = 0$, and 0.5 for scenario I and with $\gamma = 0$, 0.5 1.0 for scenario II.

Table 1 and 2 summarize the performance of the different information criteria. The positive selection rate (PSR) is defined as the ratio of identified significant predictors among all the significant predictors. The false discovery rate (FDR) is defined as the ratio of false identified predictors among all the identified predictors. In a multiple testing framework, the positive selection rates reflects the power or sensitivity of the test, and the false discovery rate reflects the error rate or selectivity of the test.

Table 1 provides the performance of different methods when $n = 200$, and $P = 30$. we observe that the strength of correlation does mildly affect the performance of different methods. But the relative comparison among different methods remain the same pattern at different correlation levels. As CL_U-AIC has less penalty on the model complexity, it always achieves higher PSR than CL_U-BIC and CL_B-BIC. Under such a modest sample size and small P setting, all the information criteria have satisfactory FDR control. With regard to the size of γ for the CL-BIC criteria, $\gamma = 0.5$ or higher seems unnecessary, and it attenuates the power. Therefore, using $\gamma = 0$ is recommended here by both CL-BIC and EBIC. The CL_B-BIC always achieve higher PSR than CL_U-BIC, demonstrating the efficiency gain by using the pairwise model rather than the univariate models. Compared to the full likelihood based EBIC, CL_B-BIC has shown PSR and FDR very close to that of EBIC. This demonstrates that under the exchangeable correlation structure, the discrepancy between the pairwise likelihood and the full likelihood is very little.

Table 2 provides the performance of different methods when $n = 200$, and $P = 1000$. With such a large number of covariates, the CL_U-AIC does not adequately control the FDR rate. It seems that CL_B-$BIC_{0.5}$ has a satisfactory performance and controls the FDR rate very well. The penalty with $\gamma = 1$ seems too harsh, and it attenuates the power. Therefore, when $P = 1000$ and $n = 200$, CL_B-$BIC_{0.5}$ is recommended. This also agrees with theorem 6 that, when $P = O(n^\kappa)$, to achieve selection consistency, it requires $\gamma \geq 1 - 1/(2\kappa)$. In this simulation setup, $P = n^{1.3}$, so $\gamma = 0.6$ is the optimal choice to ensure the consistency. The CL_B-BIC always achieves higher PSR than CL_U-BIC, suggesting the importance of incorporating correlation in the composite likelihood. The performance of CL_B-BIC is very close to that of EBIC.

4.2. Multivariate probit model

Under exactly the same setup in Section 4.1, binary correlated responses are obtained by dichotomizing the continuous multivariate normal measurements. Also, the two scenarios of $P < n$ and $P >> n$ are considered. For a multivariate probit model with many covariates, the full likelihood involves high dimensional integration and is computationally prohibitive. We thus compare the performance of the different information criteria under only composite likelihood methodology, including CL_U-AIC, CL_U-BIC, and CL_B-BIC. Results are summa-

rized in Tables 3 and 4. It is noted that even with $P = 30$, and $n = 100$, the overfitting effect of CL_U-AIC is exhibited. When $P = 1000$, the FDR of CL_U-AIC is about 50 to 70 percent, indicating an inadequate control of the error rate. The CL_B-BIC always has higher PSR than CL_U-BIC because of the advantage of using pairwise likelihood over univariate marginal likelihood. When $P = 30$, the penalty term with $\gamma = 0$ is sufficient to maintain a good FDR for CL_B-BIC. When $P = 1000$, the penalty term with $\gamma = 0.5$ is needed to control the error rate. Thus for the multivariate probit model, the CL_B-BIC is recommended for its computational feasibility and simplicity compared to the full likelihood approach, and also it clearly provides a satisfactory performance in terms of sensitivity and selectivity.

5. Concluding remarks

As a consistent model selection criterion, BIC has been widely accepted in practice. This method selects a model to achieve the balance between the model fitting and the model complexity. In contrast, AIC focuses on selecting a model with the best prediction power, which may contain unimportant predictors. Both BIC and AIC are known to be based on the so-called L_0 penalty. The LASSO method (Tibshirani, 1996) is based on a continuous L_1 penalty, allowing for the model selection among a set of infinitely many candidate models. LASSO involves a tuning parameter that needs to be determined in light of an optimal criterion. BIC, as well as AIC and cross-validation based criteria, has been widely used to determine the tuning parameter in the LASSO method or other bridge regression methods. In this sense, the proposed CL-BIC provides a feasible and rigorous tool to determine the optimal tuning parameter when the L_1 penalty is applied on composite likelihood in the high-dimensional data analysis.

Model selection is difficult when the number of parameters in the model increases with the sample size. Recently, EBIC (Chen and Chen, 2008) has been advocated to address the difficulty through adding an extra penalization term on the dimensionality of the model space. The selection consistency of the EBIC has been only established in the linear regression setting. The proposed CL-BIC may be regarded as an extension of EBIC, but it is applicable to a much broader range of likelihood or quasi-likelihood methods. The model selection consistency of CL-BIC remains true under mild regularity conditions. This is illustrated

numerically via two important statistical models. Obviously, a key advantage of the CL-BIC is that it makes the variable selection possible even if the full likelihood is not feasible to compute.

Acknowledgement

This research is supported by Natural Science and Engineering Research Council of Canada Grants held by the first author.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. Second International Symposium on Information Theory*, Ed. B. N. Petrov and F. Caski, pp 267-281. Budapest: Akademial Kiado.

Berger, J.O., Ghosh, J.K. & Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference* **112**, 241-258.

Bedrick, E.J. & Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics* **50**, 226-31.

Chakrabarti, A. & Ghosh, J.K. (2006). A generalization of BIC for the general exponential family. *Journal of Statistical Planning and Inference* **136**, 2847-2872.

Chen, J. H. & Chen, Z. H. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-71.

Cox, D. R. & Hinkey, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

Cox, D. R. & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.

Csörgö, M. & Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. New York: John Wiley & Sons.

Fearnhead, P. & Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *J. R. Statist. Soc. B* **64**, 657-80.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.* **16**, 342-355.

Heagerty, P. J. & Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Am. Statist. Assoc.* **93**, 1099-111.

Hjort, N. L. & Omre, H. (1994). Topics in spatial statistics. *Scan. J. Statist.* **21**, 289-357.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *Ann. Statist.* **35**, 1487-1511.

Konishi, S., Ando, T. & Imoto, S.(2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91**, 27-43.

Li, Y. & Lin, X. (2006). Semiparametric Normal Transformation Models for Spatially Correlated Survival Data. *J. Am. Statist. Assoc.* **101**, 591-603.

Lindsay, B. (1988). Composite likelihood methods. *Statistical inference from stochastic processes*, Ed. Prabhu, N. U., 221-239. Providence, RI: American Mathematical Society.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Parner, E. T. (2001). A composite likelihood approach to multivariate survival data. *Scand. J. Statist.* **28**, 295-302.

Rao, C. R. & Wu, Y. H. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76**, 369-74.

Renard, D., Molenberghs, G. & Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Comp. Statist. Data Anal.* **44**, 649-67.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.

Song, P.X.-K. (2007) *Correlated Data Analysis: Modeling, Analytics and Applications.* New York: Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.* **58**, 267-288.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82-6.

Tierney, L., Kass, R. E. & Kadane, J. B. (1989). Fully exponential Laplace approximation to expectations and variances of nonpositive functions. *J. Am. Statist. Assoc.* **84**, 710-716.

Varin, C. & Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519-528.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1-28.

Wang, H., Li, R. & Tsai, C.-L. (2007). Tuning parameter selector for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-68.

Wang, L. & Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. Roy. Statist. Soc., Series B* **71**, 177-190.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrika* **50**, 1-25.

Yao, Y. C. (1988). Estimating the number of change-points via Schwarz criterion. *Statist. Prob. Lett.* **6**, 181-189.

Zhao Y. & Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canad. J. Statist.* **33**, 335-356.

Appendix: Proofs

Proof of Proposition 1: Note that in the regression model with *iid* normally distributed errors, the composite likelihood of the one-dimensional marginal likelihood coincides with the full likelihood. Thus, $H_s = V_s = (X_s'X_s)/\sigma^2$. This implies that $d_s^* = d_s$.

To determine the pseudo null value $\theta_{s,0}$, we begin with

$$\begin{aligned} & 2\sigma^2 \mathrm{E}_{\theta_{T,0}}\left\{\Gamma_{T|s}(Y;\theta_{T,0},\theta_s)\right\} \\ =& -\mathrm{E}_{\theta_{T,0}}\left\{(Y-X_T\theta_{T,0})'(Y-X_T\theta_{T,0})\right\}+\mathrm{E}_{\theta_{T,0}}\left\{(Y-X_s\theta_s)'(Y-X_s\theta_s)\right\} \\ =&(X_T\theta_{T,0}-X_s\theta_s)'(X_T\theta_{T,0}-X_s\theta_s). \end{aligned} \tag{21}$$

Then its minimizer is $\theta_{s,0}=(X_s'X_s)^{-1}X_s'X_T\theta_{T,0}$. It follows that the minimum is given by

$$\begin{aligned} & \mathrm{E}_{\theta_{T,0}}\left\{\Gamma_{T|s}(Y;\theta_{T,0},\theta_{s,0})\right\}=\frac{1}{2\sigma^2}\left(X_T\theta_{T,0}-D(s)X_T\theta_{T,0}\right)'\left(X_T\theta_{T,0}-D(s)X_T\theta_{T,0}\right) \\ =&||X_T\theta_{T,0}-H(s)X_T\theta_{T,0}||_2^2/(2\sigma^2). \end{aligned} \tag{22}$$

On the other hand,

$$\begin{aligned} & \mathrm{var}_{\theta_{T,0}}\left\{\Gamma_{T|s}(Y;\theta_{T,0},\theta_{s,0})\right\}=\frac{1}{4(\sigma^2)^2}\mathrm{var}\{\epsilon'((I-D(s))X_T\theta_{T,0})\} \\ =&||X_T\theta_{T,0}-H(s)X_T\theta_{T,0}||_2^2/(4\sigma^4). \end{aligned} \tag{23}$$

Applying the above results to Assumption 4(c), it is clear that equations (9) and (10) imply equation 13.

Proof of Lemma 3: First, White's (1982) Theorem 3.2 (A.3) implies that under the regularity conditions A1-A3,

$$\frac{1}{n}\frac{\partial^2 \mathrm{cl}(\theta_s;Y)}{\partial\theta_s\partial\theta_s'}\Big|_{\hat{\theta}_s^c} \stackrel{a.s.}{\to} \mathrm{E}_{\psi_{T,0}}\left\{\frac{\partial^2 \mathrm{cl}(\theta_s;Y_1)}{\partial\theta_s\partial\theta_s'}\right\}|_{\theta_{s,0}}$$

Since $\hat{\theta}_s^c-\theta_{s,0}$ is asymptotically normally distributed with mean 0 and asymptotic covariance matrix $H_s^{-1}V_sH_s^{-1}$, by Slutsky's Theorem, $Q(s)$ asymptotically follows the same distribution as the quadratic from $\left(\hat{\theta}_s^c-\theta_{s,0}\right)'H_s\left(\hat{\theta}_s^c-\theta_{s,0}\right)$, which converges in distribution to a weighted sum of chi-square random variables, $\sum_{j=1}^{d_s}\zeta_j(s)\chi_j^2$, where $\zeta_j(s)$ are the eigenvalues of $H_s^{-1}V_s$ and χ_j^2 are *i.i.d.* with chi-square distribution with 1-degree of freedom. The proof of Lemma 3 is complete.

Proof of Lemma 4: Note that by Bonferroni Inequality, for a constant $c>0$,

$$P(\max\{Z_i, i=1,\ldots,K\}\geq c)\leq\sum_{i=1}^{K}P(Z_i\geq c).$$

In the mean while, we have

$$\begin{aligned} 1-\Phi(c) &\leq \frac{1}{c}\int_{c^2/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-z}dz \\ &= \frac{\phi(c)}{c}, \end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and density of the standard normal, respectively. This entails

$$P(\max\{Z_i, i = 1, \ldots, K\} \geq c) \leq \frac{K\phi(c)}{c} = \frac{Ke^{-c^2/2}}{\sqrt{2\pi}c}.$$

Let $c = \sqrt{2 \log K}$. Then,

$$\frac{Ke^{-c^2/2}}{\sqrt{2\pi}c} = \frac{K}{\sqrt{2\pi}K\sqrt{2 \log K}} \to 0, \text{ as } K \to \infty.$$

This leads to $\max\{Z_i, i = 1, \ldots, K\} = O_p(\sqrt{\log K})$.

Proof of Lemma 5: The space $\mathcal{S}_{d_s}$ contains a total of $\begin{pmatrix} P_n \\ d_s \end{pmatrix}$ possible submodels. For each submodel $s \in \mathcal{S}_{d_s}$, Lemma 3 entails, for large n,

$$Q(s) = \sum_{j=1}^{d_s} \zeta_j(s) Z_j^2 + o_p(1) \leq d_s^* \max\left\{Z_j^2, j = 1, \ldots, d_s\right\} + o_p(1),$$

where $Z_1, \ldots, Z_{d_s}$ are *i.i.d.* with $N(0, 1)$, and $d_s^* = \sum_{j=1}^{d_s} \zeta_j(s)$. By Lemma 4, we obtain

$$\max\left\{Q(s), s \in \mathcal{S}_{d_s}\right\} \leq d_s^* \max\left\{Z_j^2, j = 1, \ldots, d_s\right\} + o_p(1) = O_p(d_s^* \log(P_n)).$$

Table 1: Positive selection rates (PSR) and false discovery rates (FDR) on multivariate normal model with $P = 30$ and $N = 200$

β	ρ_y	rate	CL-AIC	CL_U-BIC_0	CL_U-$BIC_{0.5}$	CL_B-BIC_0	CL_B-$BIC_{0.5}$	$EBIC_0$	$EBIC_{0.5}$
β_1	0.3	(PSR)	0.878	0.739	0.655	0.911	0.875	0.914	0.877
		(FDR)	0.035	0.003	0.002	0.037	0.011	0.034	0.008
β_1	0.6	(PSR)	0.873	0.727	0.668	0.946	0.903	0.949	0.913
		(FDR)	0.026	0.002	0.002	0.053	0.014	0.055	0.017
β_2	0.3	(PSR)	0.852	0.697	0.667	0.892	0.818	0.892	0.825
		(FDR)	0.108	0.005	0.004	0.116	0.045	0.128	0.041
β_2	0.6	(PSR)	0.845	0.695	0.663	0.938	0.890	0.940	0.888
		(FDR)	0.095	0.014	0.006	0.142	0.065	0.135	0.060

Table 2: Positive selection rates (PSR) and false discovery rates (FDR) on multivariate normal model with $P = 1000$ and $N = 200$

β	ρ_y	rate	CL-AIC	CL_U-BIC_0	CL_U-$BIC_{0.5}$	CL_U-$BIC_{1.0}$	CL_B-BIC_0	CL_B-$BIC_{0.5}$	CL_B-$BIC_{1.0}$	$EBIC_0$	$EBIC_{0.5}$	$EBIC_{1.0}$
β_1	0.3	(PSR)	0.896	0.758	0.611	0.505	0.893	0.819	0.789	0.889	0.818	0.789
		(FDR)	0.472	0.035	0.002	0.000	0.439	0.039	0.012	0.378	0.037	0.011
β_1	0.6	(PSR)	0.894	0.766	0.601	0.508	0.894	0.837	0.814	0.881	0.838	0.809
		(FDR)	0.456	0.046	0.004	0.000	0.346	0.052	0.026	0.211	0.052	0.023
β_2	0.3	(PSR)	0.868	0.687	0.622	0.582	0.850	0.717	0.693	0.842	0.710	0.692
		(FDR)	0.780	0.025	0.009	0.002	0.641	0.044	0.014	0.545	0.032	0.014
β_2	0.6	(PSR)	0.873	0.688	0.612	0.575	0.847	0.728	0.703	0.815	0.722	0.702
		(FDR)	0.783	0.033	0.009	0.006	0.535	0.064	0.020	0.316	0.053	0.020

Table 3: Positive selection rates (PSR) and false discovery rates (FDR) on multivariate probit model with $P = 30$ and $N = 100$

β	ρ_y	rate	CL-AIC	CL_U-BIC_0	CL_U-$\text{BIC}_{0.5}$	CL_B-BIC_0	CL_B-$\text{BIC}_{0.5}$
β_1	0.3	(PSR)	0.846	0.710	0.670	0.768	0.682
		(FDR)	0.248	0.068	0.060	0.111	0.063
	0.6	(PSR)	0.850	0.713	0.675	0.769	0.707
		(FDR)	0.233	0.067	0.052	0.104	0.063
β_2	0.3	(PSR)	0.812	0.693	0.687	0.707	0.692
		(FDR)	0.394	0.079	0.071	0.111	0.078
	0.6	(PSR)	0.813	0.703	0.695	0.735	0.693
		(FDR)	0.363	0.089	0.065	0.130	0.069

Table 4: Positive selection rates (PSR) and false discovery rates (FDR) on multivariate probit model with $P = 1000$ and $N = 100$

β	ρ_y	rate	CL-AIC	CL_U-BIC_0	CL_U-$\text{BIC}_{0.5}$	CL_U-$\text{BIC}_{1.0}$	CL_B-BIC_0	CL_B-$\text{BIC}_{0.5}$	CL_B-$\text{BIC}_{1.0}$
β_1	0.3	(PSR)	0.790	0.766	0.593	0.393	0.782	0.647	0.475
		(FDR)	0.522	0.431	0.118	0.029	0.494	0.169	0.055
	0.6	(PSR)	0.778	0.756	0.571	0.398	0.775	0.640	0.500
		(FDR)	0.540	0.448	0.095	0.024	0.516	0.181	0.058
β_2	0.3	(PSR)	0.865	0.840	0.635	0.540	0.863	0.692	0.588
		(FDR)	0.703	0.587	0.092	0.012	0.696	0.163	0.031
	0.3	(PSR)	0.868	0.828	0.637	0.518	0.858	0.718	0.592
		(FDR)	0.711	0.590	0.087	0.012	0.678	0.167	0.036